Evaluating Whole Slide Imaging: A Working Group Opportunity

Darren Treanor MB, BSc, PhD, FRCPath[1], Brandon D. Gallas PhD[2], Marios A. Gavrielides PhD[2], and Stephen M. Hewitt MD, PhD[3],*

[1] Leeds Teaching Hospitals NHS Trust and University of Leeds
[2] Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration
[3] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health

*E-mail: *Stephen M. Hewitt - genejock@helix.nih.gov*
** Corresponding Author*

Whole slide imaging is believed to have the potential to replace the use of an optical microscope as the means of reviewing histopathology for anatomic diagnosis. Potential benefits include: improve efficiency, cost-effectiveness, and accessibility to high quality pathology review, and potentially improve diagnostic accuracy and reproducibility. Many parties - including practicing and academic pathologists, manufacturers, health system administrators, regulators, and patients - want to see the widespread adoption of safe and effective digital pathology. The widespread adoption of digital pathology has faced some obstacles, paramount of which manual interpretation has been the gold standard for over a century, but additionally economics, infrastructure, workflow, and concerns that whole slide images may not be adequate for diagnostic need. This last point is embodied by the fact that WSI has not been approved for broad use in the US due to a lack of regulatory science and data needed to assure safety and effectiveness. In an effort to advance the field of digital pathology, the authors propose an open working group focused on defining and characterizing the technical and clinical components related to digital pathology.

A whole slide image (virtual slide) is a digital reproduction of an optical image of a real physical object (a section of stained tissue on a glass slide). WSI systems, as with all digital imaging systems, capture a finite amount of information from the object. Determining what information is important within the broad scope of histo- and cyto-morphologic detail utilized by a pathologist to render a diagnosis remains unclear. Often times these issues are spoken about with language such as "Image Fidelity" or "Image Quality"; however neither term is solely appropriate or accurate. Rather the question is: how do the imaging characteristics impact the degree to which a digital image is "fit for purpose", for example, fit for some diagnostic task on a computer screen by a pathologist? In fact, the impact may be different across diagnostic tasks and may be different when image analysis is involved, as with computer-aided diagnosis. The challenge is to identify tasks that stress imaging characteristics and yield results that generalize to other tasks, as well as to design studies that compare pathologist performance completing these tasks with WSI to performance with a microscope.

Experienced pathologists can readily detect differences between the microscope image and whole slide images, and indeed between whole slide images or microscopes from different vendors. They may even cite these subjective differences as a deficiency of WSI (and as a reason for non-adoption of the technology), but it is not clear that these differences lead to differences in diagnostic performance. One reason for the lack of clarity is that the quantitative assessment of technical performance characteristics (resolution, contrast, dynamic range) throughout the WSI imaging chain (illumination, optics, scanner, image processing, transmission and display) has not been comprehensively evaluated and appreciated, despite being considered as the *first level of efficacy* to be tackled [1]. The differences between the optimal microscope and WSI are sufficiently substantial that comparisons of user performance must examine the broad scope of the activity of histopathologic diagnosis. Diagnostic performance is the second level of efficacy to be addressed [1]. If this level is approached with a clear and comprehensive understanding of the technical performance characteristics, it will be possible to link the two. At the same time, data from technical assessment can complement clinical data and may reduce the size and time (and cost) of clinical studies. Currently, it is not clear how technical performance metrics affect diagnostic performance of pathologists. On the one hand, several reports of implementation of WSI in clinical use and initial validation studies have been published, some quite large and long-term [2-4]. On the other hand, WSI systems might be generating images in which there is an inherent risk of increased diagnostic error compared to diagnoses made with the microscope. Regardless, systematic evaluation is an element of quality assurance.

Data on the performance of the pathologist in reference to technical performance is limited, at best. Apart from obvious comments about the difficulty in identifying small "objects", such as Helicobacter organisms or mitoses, or general comments about "out of focus" areas on whole-slide images, only rare qualitative allusions to the general effect of technical performance on diagnostic performance have been published. For example, difficulties interpreting dysplasia/ atypia [5]; difficulties with micrometastasis detection in one of the authors' work [6], the distinction of reactive atypia from adenocarcinoma and (less seriously) distinguishing neutrophils from eosinophils in oesophagitis [2], and discordances in the diagnosis of skin lesions attributed to difficulty in identifying eosinophils and apoptotic cells, and grading cytological atypia [7]. In all cases, the difficulties were noted to be rare and sporadic. No single causative factor was explicitly identified rather discordances were attributed by the authors to a combination of factors including scan focus, compression, colour reproduction, dynamic range, and display factors, as well as to the specific clinical problem being addressed. Gilbertson et al [5] commented that the effects of various image quality factors may be additive.

These findings have typically not been examined in quantitative experiments and resulted from early-phase studies focusing on the general application of digital pathology as a proof of concept. As such, these studies did not address the limits of the new technology, or requirements for its adoption into clinical practice. For those questions to be answered, appropriate study designs are needed. Many of the published trials and reviews of WSI are small in terms of sample size per diagnostic task [8], they often lack description of the technical characteristics of the WSI imaging chain (e.g. illumination, scanner optics, camera calibration, display calibration), and do not control for variables such as inter-scanner, inter-display, or inter-observer variability, nor differences in tissue preparation or staining protocols that might interact with image properties. Issues such as adequate sampling for specific tasks, enriching data for important population subgroups, and accounting for reader variability, need to be thoughtfully addressed. Such study designs and analyses are not trivial and have not yet been embraced in the WSI evaluation literature, but have been instrumental in the evaluation, regulatory approval, and community adoption of digital imaging systems in radiology [9-11].

This lack of knowledge and inability to address the impact of the technical characteristics in WSI leads to uncertainty with multiple consequences. We do not know with certainty whether the specifications of existing systems are adequate for general diagnostic use, whether they might be made adequate with relatively minor adjustments (e.g. to image compression levels or display settings) or additional pathologist training [12, 13], or whether digital pathology at its current technology levels may be adequate for some diagnostic tasks, but not for others. Existing users of WSI may be unaware of the technical performance levels of their device, how to maintain those levels, and what tasks they can accomplish at those levels. New users may be reluctant to take up WSI if technical or diagnostic performance is perceived as inferior to the microscope. Regulators currently do not have the data necessary to approve WSI as a primary diagnostic modality. Vendors do not have clear guidance on what constitutes "adequate" levels of technical performance for diagnostic use, relying on subjective and often conflicting pathologist opinion.

To address the issues above, we propose a working group of stakeholders (industry, clinicians, academia, and government) interested in advancing the evaluation of WSI. An overarching goal of the working group is to properly characterize WSI with systematic technical measurements and validation studies that would allow the clinical utility of digital pathology to be maximized. Much of this characterization will utilize the microscope for baseline performance expectations. The short-term objectives of the working group are:

- To form a group of interested parties
- To lay out the key technical performance metrics for WSI: gather information on the current state of the science, identify gaps in knowledge and unmet needs, and identify circumstances in which technical performance has been linked to diagnostic performance.
- To raise awareness of the issues among pathologist users, vendors, regulators, and research and healthcare funding agencies

If we are successful with the modest short-term objectives, we will consider some more ambitious long-term objectives. Possible long-term objectives are to facilitate and promote research in this area aiming to:

- Develop, standardize, and explore the range of technical performance metrics in WSI
- Design and execute experiments investigating pathologist performance as a function of image quality
- Create and disseminate methods, tools, examples, and recommendations for evaluating technical and diagnostic performance (phantoms, shared sets of slides, WSI images, protocols, study designs, analysis methods and source code)

These objectives will be further refined based on feedback and the expertise of the working group participants. We expect that the contributions of this group will make it easier for investigators to answer the key questions related to the validation of digital pathology and its adoption into clinical practice.

We invite you to join this working group (https://nciphub.org/groups/wsi_working_group), and we ask that you share this invitation to motivated and interested groups and individuals. We believe that the time and environment are ripe for this working group and have received encouragement and support from industry and government.

References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Medical Decision Making* 1991; 11: 88-94.
2. Bauer TW, Schoenfield L, Slaw RJ et al. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med* 2013; 137: 518-524.

3. Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: Digital pathology experiences 2006-2013. *Journal of Pathology Informatics* 2014; 5.

4. Evans AJ, Chetty R, Clarke BA et al. Primary frozen section diagnosis by robotic microscopy and virtual slide telepathology: the University Health Network experience. *Human pathology* 2009; 40: 1070-1081.

5. Gilbertson JR, Ho J, Anthony L et al. Primary histologic diagnosis using automated whole slide imaging: a validation study. *BMC Clinical Pathology* 2006; 6.

6. Randell R, Ambepitiya T, Mello-Thoms C et al. Effect of display resolution on time to diagnosis with virtual pathology slides in a systematic search task. *Journal of Digital Imaging* (accepted July 2014).

7. Velez N, Jukic D, Ho J. Evaluation of 2 whole-slide imaging applications in dermatopathology. *Human pathology* 2008; 39: 1341-1349.

8. Gavrielides MA, Conway C, O'Flaherty N et al. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. *Analytical Cellular Pathology* (accepted, July 2014).

9. Wagner RF, Metz CE, Campbell G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Academic Radiology* 2007; 14: 723-748.

10. Gallas BD, Chan H-P, D'Orsi CJ et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Academic Radiology* 2012; 19: 463-477.

11. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. John Wiley & Sons 2011.

12. Weinstein RS, Descour MR, Liang C et al. Telepathology overview: from concept to implementation. *Human Pathology* 2001; 32: 1283-1299.

13. Dunn BE, Choi H, Recla DL et al. Robotic surgical telepathology between the Iron Mountain and Milwaukee Department of Veterans Affairs Medical Centers: a 12-year experience. *Human Pathology* 2009; 40: 1092-1099.